# What is machine learning ?

Veille des notions

## La science des données

La science des données (data science) **est le domaine d'application des techniques d'extraction et d'analyse avancée des données.**

Plus qu'un domaine théorique, **la data science est une discipline appliquée**, dans laquelle les données facilitent la prise de décision, la planification stratégique, et plus globalement tous les processus d'une organisation, quelle que soit sa nature ou son secteur d'activité.

Elle regroupe **un large ensemble d'outils et techniques multidisciplinaires** visant à extraire de l'information exploitable de données brutes.

La data science avec ces méthodes multiples a pour objectif d'identifier des tendances, mettre en lumière des motifs des relations/corrélations entre les données.

A l'interface entre la programmation informatique et les mathématiques, la data science use d'outils tels que l'analyse prédictive ou l'optimisation en se basant sur les statistiques et l'intelligence artificielle incluant des algorithmes de machine learning par exemple.

## Le machine learning (apprentissage automatique)

Le machine learning est un domaine de l'intelligence artificielle (IA) qui se fonde sur des approches mathématiques et statistiques pour **donner aux ordinateurs la capacité d' apprendre à partir de données,** c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour cela. Il repose sur une idée fondamentale : au lieu de programmer des instructions spécifiques pour effectuer une tâche, il est possible de développer des algorithmes et des modèles qui apprennent à partir d'exemples et de données fournis.

Les machines deviennent alors capables de s'adapter à de nouvelles données et de prendre des décisions basées sur des connaissances acquises à partir de leurs expériences passées. Ainsi, les systèmes informatiques analysent les données, identifient des modèles et effectuent des prédictions.

La particularité du machine learning réside dans **sa capacité à apprendre de cet historique de données et de s'améliorer continuellement, et ce de manière totalement autonome**.

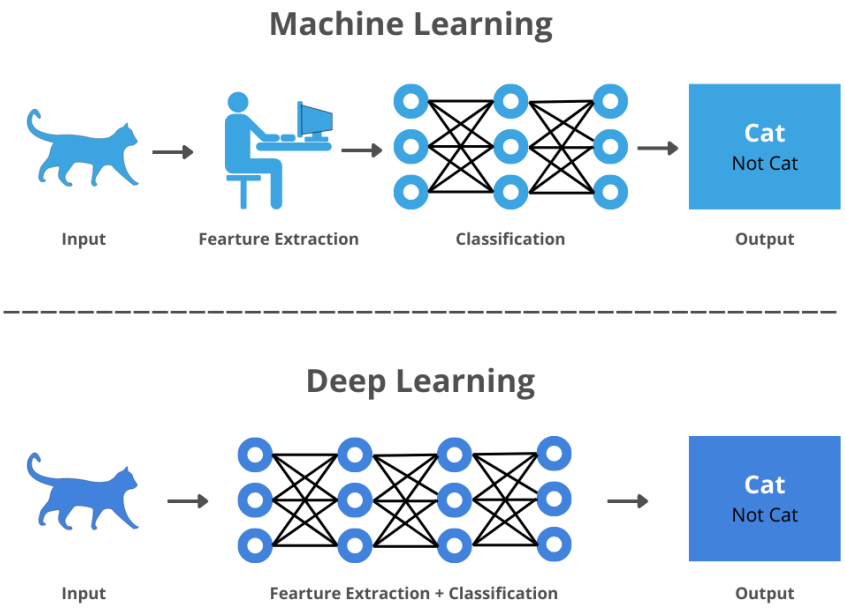
Les algorithmes de machine learning détectent des schémas ou des caractéristiques dans les données et les utilisent pour **faire des prédictions ou prendre des décisions**. Cet apprentissage peut se faire avec différente méthode :

- L'apprentissage supervisé (Supervised Learning)
- L'apprentissage non supervisé (Unsupervised Learning)
- L'apprentissage par renforcement (Reinforcement Learning)

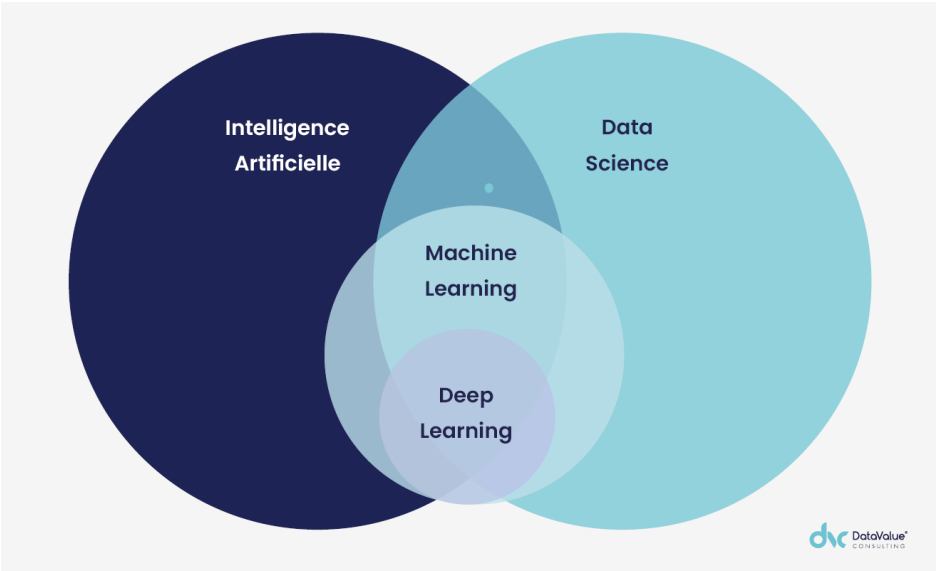
## Le deep learning (apprentissage profond)

Le deep learning est une technique de machine learning qui s'appuie sur **un réseau de neurones artificiels** s'inspirant du cerveau humain. Ce réseau est **composé de dizaines voire de centaines de « couches » de neurones**, chacune recevant et interprétant les informations de la couche précédente. Le système apprendra par exemple à reconnaître les lettres avant de s'attaquer aux mots dans un texte, ou déterminera s'il y a un visage sur une photo avant de découvrir de quelle personne il s'agit.

À chaque étape, les « mauvaises » réponses sont éliminées et renvoyées vers les niveaux en amont pour ajuster le modèle mathématique. Au fur et à mesure, le programme réorganise les informations en blocs plus complexes. Lorsque ce modèle est par la suite appliqué à d'autres cas, il est normalement capable de reconnaître un chat sans que personne ne lui ait jamais indiqué qu'il n'avait jamais appris le concept de chat. Les données de départ sont essentielles : plus le système accumule d'expériences différentes, plus il sera performant.




Alors que les algorithmes d'apprentissage automatique nécessitent généralement une correction humaine lorsqu'ils se trompent, les algorithmes d'apprentissage profond peuvent améliorer leurs résultats par la répétition, sans intervention humaine. Un algorithme d'apprentissage automatique peut apprendre à partir d'ensembles de données relativement petits, mais un algorithme d'apprentissage profond nécessite de grands ensembles de données pouvant inclure des données diverses et non structurées.



**En matière d'apprentissage automatisé, on oppose très fréquemment apprentissage supervisé et apprentissage non supervisé.**

## L'apprentissage supervisé

L'apprentissage supervisé est un type d' algorithme d'apprentissage automatique qui **apprend à partir de données étiquetées**. Les données étiquetées sont des données qui ont été étiquetées avec une réponse ou une classification correcte. L'apprentissage supervisé, comme son nom l'indique, comporte la présence d'un superviseur en tant qu'enseignant. L'apprentissage supervisé consiste à enseigner ou à entraîner la machine à l'aide de données bien étiquetées. Ce qui signifie que certaines données sont déjà étiquetées avec la bonne réponse. Après cela, la machine reçoit un nouvel ensemble d'exemples (données) afin que l'algorithme d'apprentissage supervisé analyse les données d'entraînement (ensemble d'exemples d'entraînement) et produise un résultat correct à partir des données étiquetées.

**Exemple** :
Supposons que l'on vous donne un panier rempli de différentes sortes de fruits. Maintenant, la première étape consiste à entraîner la machine avec tous les différents fruits un par un comme ceci :

- Si la forme de l'objet est arrondie et présente une dépression au sommet et est de couleur rouge, alors il sera étiqueté comme – Pomme .
- Si la forme de l'objet est un long cylindre incurvé de couleur vert-jaune, alors il sera étiqueté comme – Banane .

Supposons maintenant qu'après avoir entraîné les données, vous ayez donné un nouveau fruit séparé, par exemple une banane, dans le panier, et que vous ayez demandé de l'identifier.
Puisque la machine a déjà appris les choses des données précédentes et doit cette fois les utiliser à bon escient. Il classera d'abord le fruit avec sa forme et sa couleur, confirmera le nom du fruit comme BANANE et le placera dans la catégorie Banane. Ainsi, la machine apprend les choses à partir des données d'entraînement (panier contenant des fruits) et applique ensuite les connaissances pour tester les données (nouveaux fruits).

## L’apprentissage non supervisé

L'apprentissage non supervisé est un type d'apprentissage automatique qui apprend à partir de données non étiquetées. Cela signifie que les données n’ont aucune étiquette ou catégorie préexistante. Le but de l’apprentissage non supervisé est de découvrir des modèles et des relations dans les données sans aucune orientation explicite.

L'apprentissage non supervisé est la formation d'une machine utilisant des informations qui ne sont ni classifiées ni étiquetées et permettant à l'algorithme d'agir sur ces informations sans guidage. Ici, la tâche de la machine est de regrouper les informations non triées selon des similitudes, des modèles et des différences sans aucune formation préalable des données.

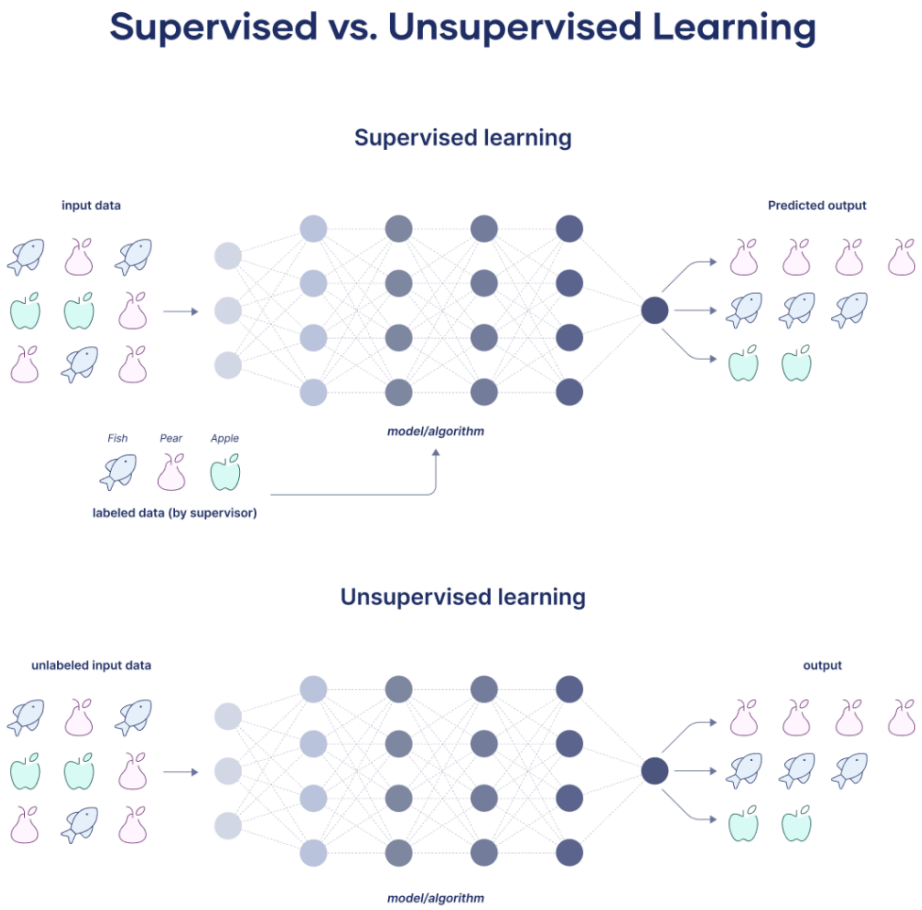
Contrairement à l’apprentissage supervisé, aucun enseignant n’est fourni, ce qui signifie qu’aucune formation ne sera dispensée à la machine. Par conséquent, la machine est limitée à trouver elle-même la structure cachée dans les données non étiquetées.

**Exemple :**
Imaginez que vous disposiez d'un modèle d'apprentissage automatique entraîné sur un vaste ensemble de données d'images non étiquetées, contenant à la fois des chiens et des chats. Le modèle n’a jamais vu d’image de chien ou de chat auparavant, et il n’a aucune étiquette ou catégorie préexistante pour ces animaux. Votre tâche consiste à utiliser l'apprentissage non supervisé pour identifier les chiens et les chats dans une nouvelle image invisible.

Par exemple , supposons qu'on lui donne une image représentant à la fois des chiens et des chats qu'il n'a jamais vus.
Ainsi, la machine n'a aucune idée des caractéristiques des chiens et des chats, nous ne pouvons donc pas la classer dans la catégorie « chiens et chats ». Mais il peut les classer selon leurs similitudes, leurs modèles et leurs différences, c'est-à-dire que nous pouvons facilement classer un tableau en deux parties. La première peut contenir toutes les photos contenant des chiens et la deuxième partie peut contenir toutes les photos contenant des chats . Ici, vous n'avez rien appris auparavant, ce qui ne signifie pas de données de formation ni d'exemples.

Cela permet au modèle de fonctionner de manière autonome pour découvrir des modèles et des informations qui n'étaient pas détectés auparavant. Il s'agit principalement de données non étiquetées.

**La distinction entre les jeux de données étiquetés et non étiquetés est la principale différence entre les deux approches.**




## La classification supervisée

La classification supervisée implique l'utilisation de données d'entraînement étiquetées pour enseigner à un modèle d'apprentissage automatique la classification de nouvelles données. Les données de formation étiquetées sont constituées d'exemples dans lesquels chaque point de données est associé à une classe ou une catégorie prédéfinie.

Le modèle d'apprentissage automatique apprend à reconnaître des modèles dans les données et à les attribuer à la classe appropriée en fonction de ces exemples prédéfinis.

## La classification non supervisée

La classification non supervisée utilise des données d'entraînement non étiquetées. L'algorithme doit identifier des modèles et des regroupements au sein des données elles-mêmes. L'algorithme recherche les similitudes et les différences entre les points de données et les regroupe en fonction de ces similitudes.

## La régression

La régression sert à trouver la relation d'une variable par rapport à une ou plusieurs autres. Dans l'apprentissage automatique, le but de la régression est **d'estimer une valeur (numérique) de sortie à partir des valeurs d'un ensemble de caractéristiques en entrée**.

**Exemple :**

Supposons que vous souhaitez prédire le prix d'une maison en fonction de sa superficie, du nombre de chambres et de son emplacement. Vous pouvez utiliser un modèle de régression pour apprendre la relation entre ces variables et le prix de la maison. Le modèle apprendra à pondérer chaque variable pour obtenir une prédiction précise du prix.

**Types de modèles de régression :**

- Régression linéaire : Ce modèle suppose une relation linéaire entre les variables prédictives et la variable cible.
- Régression polynomiale : Ce modèle suppose une relation polynomiale entre les variables prédictives et la variable cible.
- Régression logistique : Ce modèle est utilisé pour la classification binaire, où la variable cible peut prendre deux valeurs (par exemple, 0 ou 1).

## La cross validation (validation croisée)

La cross validation ou validation croisée est une méthode de Machine Learning qui permet d'évaluer les performances des modèles d'apprentissage automatique. Lorsqu'on entraîne un modèle sur des données étiquetées, on émet l'hypothèse qu'il doit également fonctionner sur de nouvelles données. Une confirmation supplémentaire sera tout de même nécessaire pour s'assurer de l'exactitude ou non de ses prédictions. La validation croisée permet justement de vérifier si cette hypothèse est valide, ou non. On pourra ensuite choisir l'algorithme de Machine Learning approprié pour effectuer une tâche précise.

## Les données d'entraînement, les données de test et de validation

**Données d'entraînement :**
Les données d'entraînement sont un ensemble de données utilisé pour apprendre un modèle d'apprentissage automatique. Elles contiennent des exemples de la tâche que le modèle doit apprendre à effectuer. Le modèle analyse les données d'entraînement et apprend à identifier les patterns et les relations entre les variables pour pouvoir ensuite effectuer des prédictions sur de nouvelles données.

**Données de validation :**
Les données de validation sont un ensemble de données utilisé pour ajuster les hyper paramètres d'un modèle d'apprentissage automatique. Elles se situent entre les données d'entraînement et les données de test en termes d'utilisation. Elles peuvent être utilisées pour comparer différents modèles et choisir celui qui s'adapte le mieux aux données.

**Données de test :**
Les données de test sont un ensemble de données utilisé pour évaluer la performance d'un modèle d'apprentissage automatique sur des données invisibles. Elles ne doivent pas être utilisées pour entraîner le modèle. Le modèle est testé sur les données de test pour mesurer sa capacité à généraliser aux données qu'il n'a jamais vues auparavant.

**Différences clés :**

- Utilisation: Les données d'entraînement sont utilisées pour apprendre le modèle, les données de test sont utilisées pour évaluer la performance du modèle et les données de validation sont utilisées pour ajuster les hyperparamètres du modèle.
- Visibilité pour le modèle: Les données d'entraînement et les données de validation sont visibles pour le modèle pendant l'apprentissage, tandis que les données de test ne le sont pas.
- Taille: Les données d'entraînement sont généralement le plus grand ensemble de données, suivies des données de validation et des données de test.

**Exemple :**
Supposons que vous souhaitez entraîner un modèle pour prédire le prix d'une maison. Vous pouvez diviser votre ensemble de données en trois parties :

- Données d'entraînement : 70% des données
- Données de validation : 15% des données
- Données de test : 15% des données

Le modèle sera ensuite entraîné sur les données d'entraînement et ses hyperparamètres seront ajustés sur les données de validation. La performance finale du modèle sera évaluée sur les données de test.

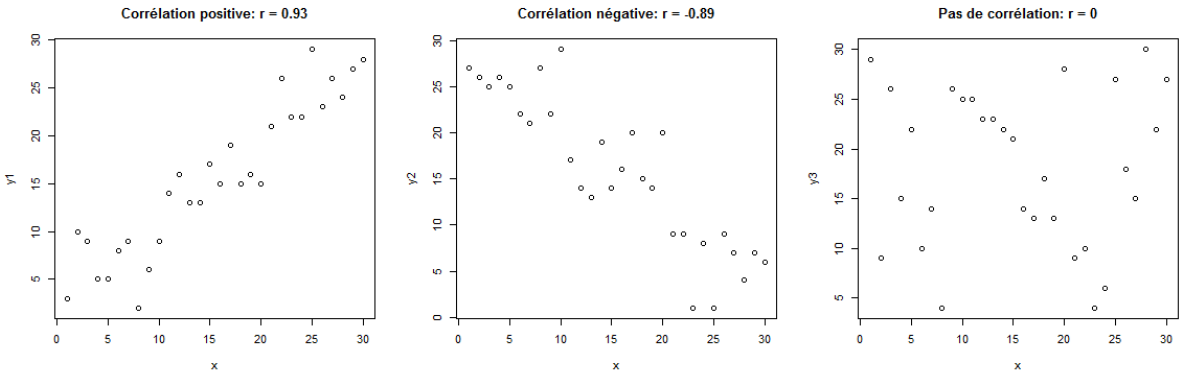
## Le coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson, également connu sous le nom de coefficient r, est une mesure statistique qui définit la force de la relation entre deux variables et leur association l'une avec l'autre.

En termes simples, le coefficient de corrélation de Pearson détermine tout changement dans une variable qui est influencée par l'autre variable liée. Le coefficient de corrélation de Pearson est influencé par le concept de covariance, ce qui en fait une meilleure méthode pour déterminer la relation et l'interdépendance entre les deux variables.

Le coefficient de corrélation varie entre -1 et +1, 0 reflétant une relation nulle entre les deux variables, une valeur négative (corrélation négative) signifiant que lorsqu'une des variable augmente, l'autre diminue ; tandis qu'une valeur positive (corrélation positive)

indique que les deux variables varient ensemble dans le même sens. Voici des exemples illustrant les 3 situations:




## Une fonction de coût

Dans le domaine de l'intelligence artificielle, la fonction de perte ou de coût est la quantification de l'écart entre les prévisions du modèle et les observations réelles du jeu de données utilisé pendant l'entraînement.

La phase d'entraînement vise à trouver les paramètres du modèle qui permettront de minimiser cette fonction.

## La descente de gradient

La descente de gradient est un algorithme d'optimisation couramment utilisé pour entraîner des modèles d'apprentissage automatique et des réseaux neuronaux. Les données d'entraînement aident ces modèles à apprendre au fil du temps, et la fonction de coût dans la descente de gradient agit spécifiquement comme un baromètre, évaluant sa précision à chaque itération des mises à jour des paramètres. Jusqu'à ce que la fonction soit proche de zéro ou égale à zéro, le modèle continue à ajuster ses paramètres pour obtenir l'erreur la plus faible possible. Une fois que les modèles d'apprentissage automatique sont optimisés et acquièrent une précision satisfaisante, ils peuvent se révéler des outils puissants pour l'intelligence artificielle (IA) et les applications informatiques.

src : https://www.lepont-learning.com/fr/data-science-definition-enjeux/
https://www.salesforce.com/fr/resources/definition/machine-learning/
https://zaion.ai/ressources/actualites/quest-ce-que-le-machine-learning/
https://www.hpe.com/fr/fr/what-is/deep-learning.html
https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/
https://www.coursera.org/articles/ai-vs-deep-learning-vs-machine-learning-beginners-guide
https://www.alteryx.com/fr/glossary/supervised-vs-unsupervised-learning
https://forestryblog.com/differences-between-supervised-classification-unsupervised-classification/
https://projeduc.github.io/intro_apprentissage_automatique/regression.html
https://www.jedha.co/formation-ia/cross-validation#:~:text=La%20cross%20validation%20ou%20validation,fonctionner%20sur%20de%20nouvelles%20donn%C3%A9es.
https://gemini.google.com/app
https://www.voxco.com/fr/blog/coefficient-de-correlation-de-pearson/#:~:text=Le%20coefficient%20de%20corr%C3%A9lation%20de%20Pearson%2C%20%C3%A9galement%20connu%20sous%20le,'une%20avec%20l'autre.
https://www.cnil.fr/fr/definition/fonction-de-perte-ou-de-cout-loss-function
https://www.ibm.com/fr-fr/topics/gradient-descent